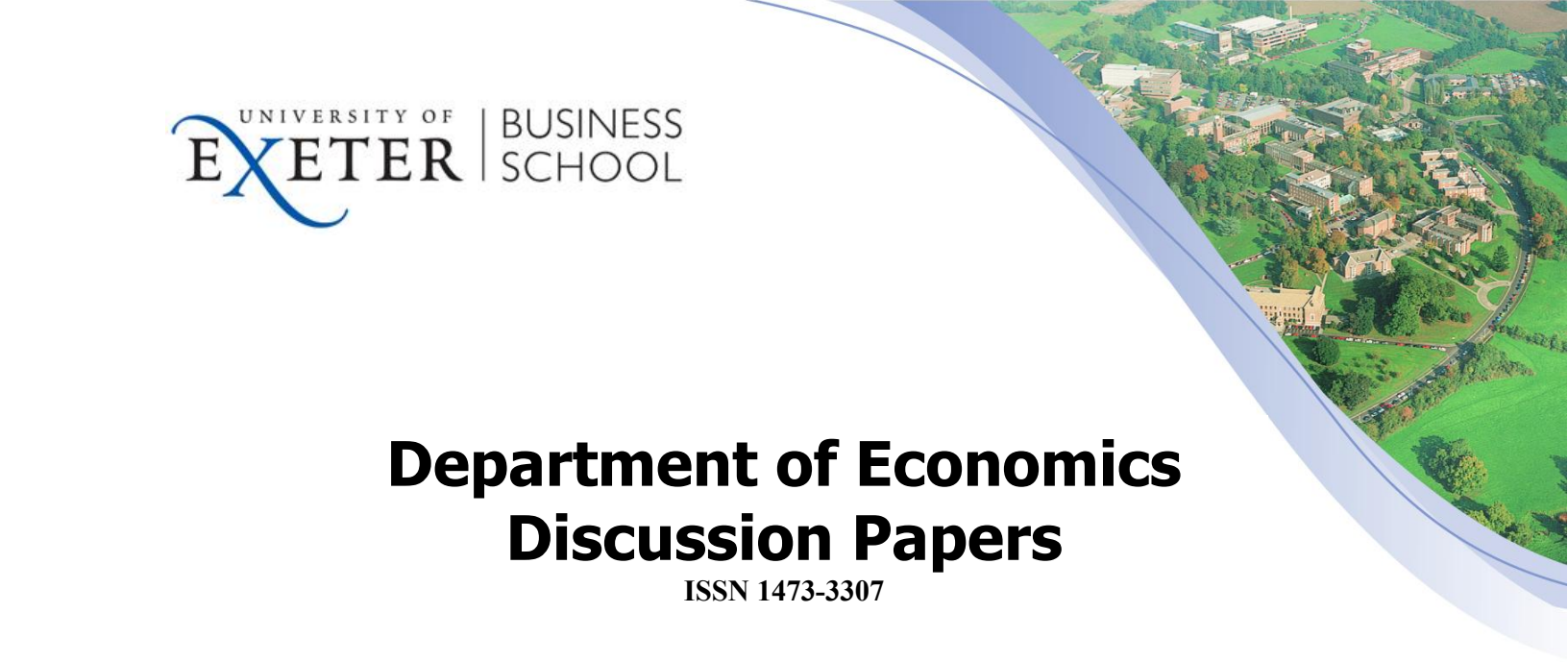


Department of Economics Discussion Papers

ISSN 1473-3307

Leader Gender, Participation, and the Quality of Contributions in Groups

Brit Grosskopf and Yangfei Lin

Paper number 26/01

RePEc page: https://ideas.repec.org/s/exe/wpaper.html

Leader Gender, Participation, and the Quality of Contributions in Groups*

Brit Grosskopf
University of Exeter

Yangfei Lin
Zhejiang University

January 8, 2026

Abstract

In a controlled laboratory experiment, we study how leader gender affects both the willingness of group members to contribute ideas and the informational quality of those contributions. Participants are randomly assigned to groups of three, consisting of one merit-based leader and two group members. Group members submit answers to a general knowledge task and report their willingness—on a scale from one to five—to have their answer selected as the group response. In the baseline condition, gender information is not revealed; in the treatment condition, the gender of the leader and group members is disclosed. We find that men become significantly more willing to contribute when led by a female leader, while women's willingness does not depend on leader gender. However, this increase in male participation is accompanied by a decline in the accuracy threshold at which men are willing to step forward. In contrast, women raise their accuracy threshold under female leadership, contributing only when they are highly confident in the correctness of their answers. As a result, conditional on stating the highest willingness level, men are substantially less accurate under female leadership, whereas women are more accurate. We refer to this asymmetric pattern as a *sisterhood effect*. We find no corresponding brotherhood effect: men do not exhibit higher conditional accuracy under male leadership relative to the no-gender benchmark. On the selection side, leaders strongly weight stated willingness

*Corresponding author: Yangfei Lin, yangfeilin91@gmail.com. We thank Katherine B. Coffman and Helena Fornwagner and the audiences at the 4th Cardiff Economic Theory and Experimental Economics workshop, the 4th Newcastle Experimental Economics workshop, the MPI in Bonn and the University of Heidelberg for valuable comments. Preliminary results from the experimental study were discussed in the PhD thesis of Yangfei Lin.

when choosing whose answer represents the group. When gender is revealed, female leaders are more likely to select female group members, whereas male leaders show no systematic gender-based selection. This selection behaviour mitigates the lower quality of highly willing male contributions under female leadership and preserves group performance. Overall, leader gender shapes collective decision-making not by altering underlying ability, but by changing how private knowledge is translated into expressed willingness and how that willingness is filtered into group choices.

1 Introduction

The contribution of knowledge is central to effective group decision-making. By aggregating information and knowledge from multiple individuals, groups can improve the quality of collective outcomes (Charness & Chen, 2020; Evans et al., 2024; Hamilton et al., 2003). Leadership plays a key role in this process, as leaders influence not only which ideas are selected, but also who is willing to step forward and contribute their knowledge. While a large literature studies how female leaders affect organizational performance (e.g., Ahern and Dittmar, 2012; Carter et al., 2003; Faccio et al., 2016; Matsa and Miller, 2011, 2013; Zhang, 2020) and follower behaviour (e.g., Beaman et al., 2012; Gangadharan et al., 2016; Grossman et al., 2015; Karpowitz et al., 2024), comparatively little is known about how gender diversity in leadership shapes the *provision* and *selection* of ideas within groups. In particular, we lack evidence on how leader gender affects who speaks up, how confident contributors must be to do so, and how leaders respond to these contributions.

We address these questions using a controlled laboratory experiment that builds on Coffman (2014) but differs in several important respects. In Coffman (2014), individuals decide whether to enter a competitive environment, and outcomes are mechanically determined once entry occurs; there is no leader who actively selects among contributors. By contrast, we introduce a merit-based leader who chooses which group member represents the group for each decision. This design allows us to separate the *supply of ideas*—who is willing to step forward—from the *selection of ideas*—whose contributions are ultimately chosen. Crucially, it also allows us to study how stated willingness maps into contribution quality by examining accuracy conditional on willingness, a margin that is not observable in settings without an active selection stage.

Participants first complete an individual task measuring ability across gender-

typed domains. The top one-third of performers are selected as leaders, a process unknown to participants during the individual task. Leadership is therefore merit-based and orthogonal to gender differences in competitiveness or preferences for leadership roles (Niederle & Vesterlund, 2007). Leaders and group members are then randomly assigned to groups of three. In the group stage, members answer new general knowledge questions and report their willingness—on a scale from one to five—to have their answer selected as the group response. Leaders observe past performance and stated willingness and select one member to represent the group for each question.

We vary whether gender information is revealed. In the baseline condition, participants have no information about each other's gender. In the treatment condition, leader gender is disclosed to group members and group member gender is disclosed to the leader. This design isolates the causal effect of leader gender on both the supply and selection of ideas while holding constant information about ability and past performance.

Our results reveal a pronounced asymmetry in how men and women respond to leader gender. When gender is known, men become more willing to contribute under female leadership than under male leadership or when leader gender is unknown, but do so at a lower accuracy threshold. Women's willingness does not increase under female leadership; however, conditional on stating maximum willingness, women's contributions are substantially more accurate under female leaders. We refer to this asymmetric pattern as a *sisterhood effect*. We find no corresponding brotherhood effect: men do not exhibit higher conditional accuracy under male leadership relative to the no-gender benchmark.

On the selection side, leaders strongly weight stated willingness when choosing whose answer represents the group. When gender is revealed, female leaders are more likely to select female contributors, whereas male leaders show no systematic gender-based selection. This selection behaviour mitigates the lower quality of highly willing male contributions under female leadership and preserves group performance. A simple counterfactual selection rule based on these patterns improves performance in male-led groups but yields no gains in female-led groups.

Our findings contribute to a growing literature documenting challenges faced by female leaders, including reduced support and resistance from subordinates (e.g., Abel, 2024; Ayalew et al., 2021; Boring, 2017; Chakraborty and Serra, 2024; Grossman et al., 2019). We identify a novel mechanism through which leader gender shapes group decision-making: by altering the accuracy threshold at which individ-

uals are willing to contribute and by affecting how leaders filter contributions. More broadly, our results show that leader gender can influence collective outcomes not only through participation rates but through systematic changes in how information is expressed and aggregated.

The remainder of the paper is organized as follows. Section 2 describes the experimental design. Section 3 presents the hypotheses. Section 4 reports the results, and Section 5 concludes.

2 Experimental Design

We employ a between-subject design with two conditions: a *Baseline* condition in which participants have no information about each other's gender, and a *Treatment* condition in which gender is implicitly revealed. The manipulation is implemented by recruiting a gender-balanced subject pool and assigning male and female participants to different sides of the laboratory, allowing participants to infer gender through location without any explicit announcement.[1] This approach makes gender salient while minimizing experimenter demand effects.

Table 1: Description of between-subject design

	Baseline	**Treatment**
Leader	No gender information about group members	Gender information about group members (implicitly)
Group members	No gender information about the leader	Gender information about the leader (implicitly); No gender information about the other group member

The experiment consists of four parts (Parts A–D), followed by a non-incentivized elicitation of gender stereotypes and a demographic questionnaire. Participants complete the four parts in fixed order. Part A contains two incentivized tasks, while Parts B–D each contain one incentivized task. At the end of the experiment, one of the five incentivized tasks is randomly selected for payment. Participants re-

[1]Participants selected tokens from different buckets corresponding to different seating areas, without being informed that these buckets were gender-specific.

ceive £0.50 per point earned in the selected task, in addition to a £5 show-up fee. Instructions are provided separately at the beginning of each part.

Figure 1 summarizes the experimental structure. Part A measures individual ability and beliefs about own performance. Part B elicits group members' willingness to step forward and contribute knowledge, as well as leaders' selection decisions. Part C elicits beliefs about group performance. Part D measures individual risk preferences. Crucially, gender information is only revealed in Part B in the *Treatment* condition, allowing us to isolate the effect of leader gender on contribution and selection behaviour.

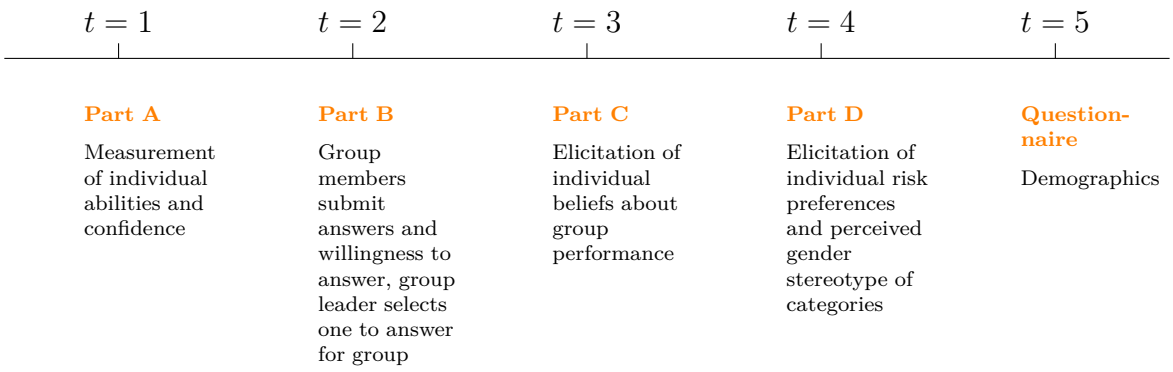


Figure 1: Timeline of the experiment

2.1 Part A

Part A consists of two incentivised tasks. In the first task, participants independently answer multiple-choice questions in six categories — Art, Entertainment, Environmental Science, Geography, History, and Sports — with five questions per category. This task provides a measure of individual ability in gender-typed domains. Following prior work (e.g., Chen and Houser, 2019; Coffman, 2014), Art and Entertainment are considered female-typed categories, while the remaining categories are considered male-typed. Participants earn one point for each correct answer if this task is selected for payment.

In the second task, participants report their subjective probability that each of their answers is correct. To elicit beliefs, we employ the incentive-compatible mechanism introduced by Karni (2009) and used in Coffman (2014) and Möbius et

al. (2022). Participants are informed that, for each question, a robot with a randomly drawn accuracy level between 1 and 100 percent may answer on their behalf. If the participant's stated belief exceeds the robot's accuracy, the participant's answer is submitted; otherwise, the robot's answer is used. If this task is selected for payment, participants receive one point for each correct answer regardless of whether it comes from the participant or the robot.

2.2 Part B

Participants are randomly assigned to groups of three, consisting of one leader and two group members. Leaders are selected based on performance in Part A: participants whose scores place them in the top third of the distribution are designated as leaders. Importantly, participants are not informed of this selection rule before or during Part A, ensuring that leader assignment is orthogonal to gender differences in competitiveness or preferences for leadership (e.g., Niederle and Vesterlund, 2007). Leaders receive a fixed payment of five points.

Group members answer a new set of multiple-choice questions, again spanning the six categories, and report their willingness to have their answer chosen as the group response. Willingness is reported on a five-point scale ranging from weak willingness (1) to strong willingness (5). After observing both group members' answers, willingness reports, and their Part A performance by category, the leader selects one answer to represent the group for each question. Leaders do not observe the content of the question itself.

If Part B is selected for payment, each correct group answer earns one point for all group members. In addition, the group member whose answer is selected receives 0.1 points regardless of correctness. This small bonus provides an incentive to submit answers while preserving strong incentives for leaders to select correct answers and for group members to truthfully reveal their willingness to contribute.

In the *Baseline* condition, neither leaders nor group members receive any information about the gender of other participants. In the *Treatment* condition, gender is implicitly revealed through seating location: men and women are seated on different sides of the laboratory, marked with distinct symbols. Participants are informed of the seating location of their leader and leaders are informed of the seating location of their group members, allowing them to infer gender without explicit labelling. Group members are never informed of the gender of the other group member, ensuring that the manipulation isolates the effect of leader gender. This design isolates the effect

of leader gender on group members' willingness to contribute, holding constant any information about peer gender.

2.3 Part C

In Part C, participants report their beliefs about group performance by estimating the number of questions their group answered correctly in Part B, by category. If this task is selected for payment, each correct estimate receives 2.5 points. Incorrect estimates are not penalized.

2.4 Part D

Part D elicits individual risk preferences. Each participant receives an endowment of five points and chooses between a safe option yielding zero points and a risky option. The risky option yields one point if a randomly drawn number is below a given threshold and results in a loss of 0.25 points otherwise. Participants make this choice for eight thresholds ranging from 20 to 90 in increments of ten.

3 Hypotheses

In our experiment, the role of a leader is introduced before group members declare their willingness to contribute their knowledge. As a result, the interpretation of willingness to answer for the group differs from that in previous studies (e.g., Chen and Houser, 2019; Coffman, 2014), where willingness to answer is often interpreted as a willingness to compete or to lead. In our setting, leadership is fixed and leaders make decisions on behalf of the group by selecting one of the two group members to represent the group. Because group members neither compete for leadership nor bear individual costs from stepping forward, willingness to answer can be interpreted as a willingness to step forward and contribute one's knowledge rather than a desire to lead or compete.

The hypotheses below were preregistered prior to data collection. They reflect a natural benchmark implied by theories of in-group bias and homophily, which predict symmetric same-gender effects for male and female leaders. In particular, we base our hypotheses on social identity theory (e.g., Haslam et al., 2015; Hogg and Van Knippenberg, 2003; Hogg et al., 2012; Tajfel et al., 2001), which emphasizes

that shared social identity between leaders and followers can influence both followers' willingness to contribute and leaders' selection decisions. Deviations from this benchmark are informative about how leader gender shapes both the supply of ideas and their selection.

Hypothesis 1: *A group member's willingness to step forward and contribute knowledge will be greater when paired with a leader of the same gender.*

Hypothesis 2: *A leader will prefer a group member of the same gender and select them to answer for the group.*

4 Results

4.1 Summary statistics

We conducted the laboratory experiment at the Finance and Economics Experimental Laboratory at Exeter (FEELE).[2] The participants were students of the University of Exeter. The experiment was conducted from February 2022 to March 2022. We ran 13 computerized sessions programmed in zTree (Fischbacher, 2007). We recruited 135 participants for the experiment from the FEELE subject pool. 63% participants were white, 26% Asian, 7% mixed, 1% black with the rest self-identifying as other. We aimed to recruit gender-balanced samples in both *Baseline* and *Treatment*. However, we ended up with 51 participants in *Baseline*: 23 males and 28 females; and 84 participants in the *Treatment*: 42 males and 42 females due to differential show-up rates. As can be seen from Table 2 men score higher in the individual general knowledge task than women. There are no other differences between the sexes.

In the second part of the experiment, participants were randomly assigned into groups of three, with one leader in each group, which means that there were 34 (56) group members in *Baseline* (*Treatment*), respectively. As a result of the better performance of men in Part A, we have slightly more male leaders than female ones, in particular in *Treatment* where the difference in performance was more pronounced (15.45 vs. 13.88, $p = 0.01$) in contrast to *Baseline* (14.78 vs. 13.39, $p = 0.09$). There were 9 male leaders and 8 female leaders in *Baseline* and 18 male leaders and 10

[2]This experiment was pre-registered at AEA (AEARCTR-0008914). Ethical approval was obtained from the Ethics Committee of the University of Exeter Business School (eUEBS002788).

Table 2: Summary statistics – all participants (mean values)

	Men ($N = 65$)	Women ($N = 70$)	p-value (M = W)
Part A score	15.22	13.67	< 0.01
Age	20.22	20.43	0.59
Payment	11.35	11.23	0.81
Risk	5.51	5.19	0.26

Notes: p-values are from the Fisher–Pitman permutation test.

female leaders in the *Treatment*.[3] To address Hypothesis 1, we will focus on the behaviour of the group members in the following.

Table 3 summarizes the performance of group members in the individual task of Part A. Note that there are no gender differences between the correct answers in the different categories. This is due to the fact that the best participants have been selected to be leaders at this point.[4]

Table 3: Performance of Group Members in Part A (Mean Scores)

	Men ($N = 38$)	Women ($N = 52$)	Total ($N = 90$)	p-value (M = W)
Art	1.89	2.10	2.01	0.45
Geography	2.32	2.06	2.17	0.34
Sports	2.58	2.37	2.46	0.29
Entertainment	1.13	1.13	1.13	1.00
Environmental Science	3.76	3.44	3.58	0.19
History	1.55	1.38	1.46	0.49

Notes: The table reports mean scores by subject and gender. The p-values are from Fisher–Pitman permutation tests of equality of means between men and women. Expected scores are reported in the Appendix, Table 13.

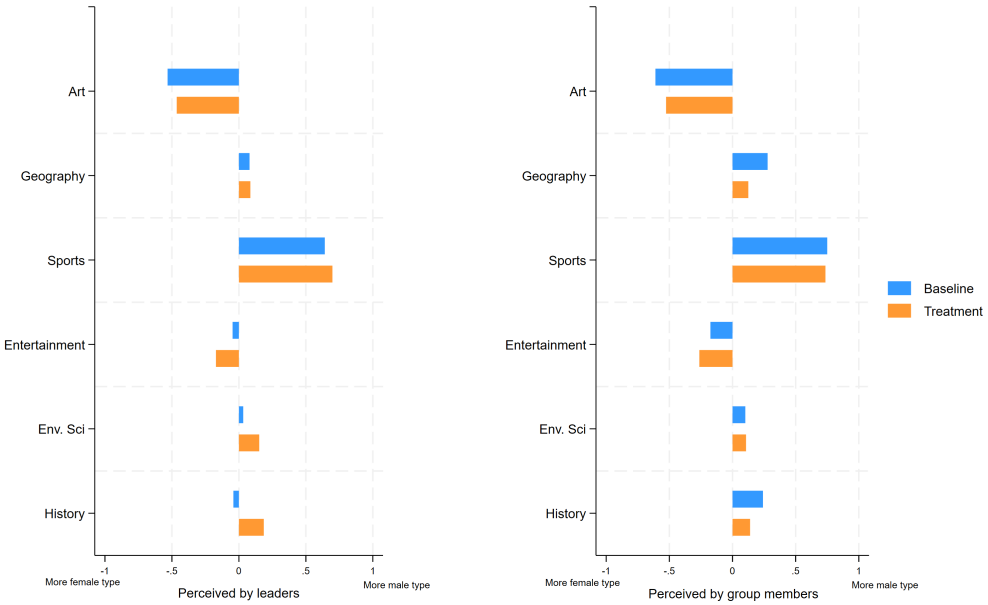
At the end of the experiment, participants were asked if they think men or

[3]We tested for performance differences between male and female leaders. There is no difference ($p = 0.25$, Fisher Pitman permutation test).

[4]Looking at the performance of all subjects, including those that became leaders, we find that men are better at environmental science, geography and sports (see Table 12 in the Appendix).

women are better in the different categories. This perception of gender stereotypes was generated for each category by choosing a number from -1 to 1, where -1 stands for "Women know more", 1 refers to "Men know more" and 0 means there are no perceived gender differences.

Figure 2: Average perceived gender-type



Notes: We test whether these averages of perceived gender stereotypes are significantly different from 0 (no gender stereotyping) using the Fisher-Pitman permutation test. We first take a look at group members. In *Baseline*, significant differences are found for all categories ($p < 0.01$) except for Environmental Science ($p = 0.09$). In *Treatment* this is also the case with Art ($p < 0.01$), Geography ($p = 0.05$), Sports ($p < 0.01$), Entertainment ($p < 0.01$) and History ($p = 0.03$) except Environmental Science ($p = 0.07$). In terms of leaders, perceived gender stereotypes are only found in Art and Sports in the *Baseline*, $p < 0.01$). Leaders in the *Treatment* consider Art ($p < 0.01$), Sports ($p < 0.01$), Entertainment ($p < 0.01$), Environmental Science ($p = 0.02$) and History ($p = 0.03$) to be gender typed, but not Geography ($p = 0.31$).

The results in Figure 2 indicate that group members perceive *Art* and *Sports* as the most gender-stereotyped, with *Art* strongly perceived to be a female-stereotyped domain and *Sports* a male-stereotyped domain. *Entertainment* is perceived as a slightly female-stereotyped domain, and the remaining categories are perceived as

male-stereotyped domains. This is consistent with the findings of Coffman (2014) and Chen and Houser (2019). *Environmental Science* does not seem to be gender-typed (only marginally so at $p = 0.09$ and $p = 0.07$ in *Baseline* and *Treatment* respectively, Fisher-Pitman permutation test). In the next section, we present results about whether the gender of the leader will change the group member's willingness to answer for the group using regression analysis.

4.2 A group member's willingness to answer

Observation 1 (Leader gender and willingness). *When leader gender is revealed, male group members state a higher willingness to contribute when paired with a female leader. Female group members' stated willingness does not vary with leader gender.*

Support Table 4 reports the results of Tobit regressions that predict the willingness of group members to answer for their groups. The dependent variable is the willingness of a group member to answer for the group at the question level, by construction censored to lie between 1 and 5.[5] Table 4 gives a first insight into treatment differences taking into account the gender of the leader, the group member's ability, and the group member's risk preferences. We will include more variables in later regressions. The first column of Table 4 presents the independent variables. "Diff leader" is a dummy that indicates whether the group member was paired with a leader of the other gender. "Ans *QiCorr*" is a proxy for the question-specific ability equal to 1 if the group member correctly answered question i in Part B; otherwise, equal to 0. "Risk" measures risk preferences, with higher values indicating higher levels of risk-loving.[6] "Female" is a gender dummy, which equals 1 if the participant is female.

The results in Table 4 provide a simple validation check for the design. A different gender leader should not have an impact on the willingness to answer in *Baseline*, where participants did not have gender information about others in their

[5]We find similar patterns in the results of ordered Probit regressions predicting willingness to contribute, which are presented in the Appendix, Table 19.

[6]In our experimental design, a group member was rewarded if they were chosen by the leader to answer for the group. This could create additional motivations for a group member to state a higher willingness to answer. For example, a group member believes that the group is unlikely to perform well and thus will state a high willingness so that at least they will receive the reward from being chosen by the leader. Neither men nor women seem to think that their groups will do particularly poorly, see Table 15 in the Appendix.

Table 4: Tobit Estimates Predicting Willingness to Answer with Risk Preferences

	Men		Women		Pooled	
	Baseline	Treatment	Baseline	Treatment	Baseline	Treatment
Diff leader	-0.219	1.176**	0.230	-0.014	-0.216	1.072**
	(0.529)	(0.585)	(0.576)	(0.405)	(0.535)	(0.499)
Ans $QiCorr$	1.975***	2.151***	2.045***	2.158***	2.024***	2.151***
	(0.269)	(0.314)	(0.322)	(0.240)	(0.219)	(0.190)
Risk	0.282*	0.399**	0.015	0.291**	0.118	0.333***
	(0.163)	(0.166)	(0.226)	(0.114)	(0.149)	(0.096)
Diff leader					0.476	-1.127*
× Female					(0.779)	(0.677)
Constant	0.685	-0.514	1.504	0.340	1.552*	-0.108
	(0.820)	(1.081)	(1.154)	(0.566)	(0.817)	(0.659)
Observations	420	720	600	960	1,020	1,680

Notes: The sum of the coefficients on "Diff leader" and "Diff leader × Female" equals zero (Wald test, $p = 0.89$). Appendix Table 20 reports Romano–Wolf adjusted p-values controlling for multiple hypothesis testing (Clarke et al., 2020; Romano & Wolf, 2005, 2016). Results remain significant after correcting for multiple hypothesis testing. Standard errors clustered at the individual level are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

group. Columns (2), (4) and (6) report Tobit regressions for *Baseline* in which "Diff leader" has no significant effect on group members' willingness to answer. In contrast, the coefficients of "Diff leader" are significant and positive in columns (3) and (7), indicating that a group member is more willing to answer for the group when paired with a leader of a different gender. These findings are driven by the behaviour of men. Women do not increase their willingness to contribute when matched with a male leader compared to being matched with a female leader. This can be seen in the individual regressions and in the pooled one. The positive coefficient of "Diff leader" is directly offset by the negative coefficient of "Female×Diff Leader" (Wald test, $p = 0.89$) in the pooled regression.

Observation 2 (Role of knowledge). *Group members who answer a question correctly are more likely to state a higher willingness to contribute that answer to the group.*

Support Intuitively, if a group member thinks they know the answer to a question, a higher willingness to answer for the group should be stated. As shown in Table 4, "Ans *QiCorr*" significantly increases male as well as female group member's willingness to answer in both *Baseline* as well as *Treatment*. The magnitude of the effect of "Ans *QiCorr*" is sizable. For example, if a group member knows the answer to a question, the stated willingness is likely to move from, say, "neutral" to "moderately strong" on the scale. It makes a group member about 'one point' more likely to want to step forward and answer for the group.

In addition to the effect of the gender of the leader, and the knowledge effect we also observe an effect of risk attitudes.

Observation 3 (Risk attitudes) *When gender is known, both men and women are more willing to step forward when they are more risk loving.*

Support As shown in Table 4, "Risk" is significantly positive for both men and women in *Treatment*.

In addition to the above mentioned observations, we also observe behaviour by both women and men that relates to gender stereotypes. The fourth observation is presented in three parts.

Observation 4a (Domain specific female gender stereotype threat) *Women are less willing to contribute answers in male-type domains regardless of whether gender is known or not.*

Support To address gender stereotypes, we run additional Tobit regressions where we include additional explanatory variables. "PartA score" is a performance

measure by category. It measures the group member's score in the specific category of the individual task in Part A. For example, if question i is in *Art*, then "PartA score" refers to that group member's score in *Art* in the individual task of Part A. This performance measure is in addition to "Ans Q_iCorr" which measures performance of the specific question i in Part B. We also include the perceived maleness of the category. "Maleness" is the average maleness (as perceived by the group members) of the category from which question i was drawn.

From Table 5 we see that the coefficient "Maleness" is significantly different from zero and negative for the female only regressions. It shows that women are less willing to contribute in male-type domains regardless of whether gender is known or not. Although this has been observed before (Kanter, 1977; Chen and Houser, 2019; Bordalo et al., 2016), it is remarkable to see that even without making gender salient, this effect is there, i.e., women do not need to be primed of their gender to feel threatened. The positive coefficient of "Maleness" in the pooled regression is more than offset by the negative coefficient of "Female×Maleness". Men respond differently when gender information is provided.

Observation 4b (Domain specific male stereotype posture effect) *Men show a reverse domain specific stereotype threat. They are more willing to contribute answers in male-type domains when gender is known.*

Support Table 5 shows that the coefficient "Maleness" is significantly different from zero and positive for men in *Treatment* indicating they are more willing to contribute their answers in male-type domains when gender is known. This shows that the gender stereotype effect is more salient when information about the gender of the leader/group members is given since it is more likely to activate gender stereotypes. However, previous results have mostly focused on men in female-type domains and women in male-type domains.

Observation 4c (Gender primed female stereotype threat) *Women are more willing to contribute in Part B when they have done well in Part A but only when not primed of their gender and that of their leader.*

Support Table 5 shows that the coefficient "PartAscore" is significant for female regressions in *Baseline* but not in *Treatment*.

To summarize, we find that the willingness to step forward and answer for the group is determined by four things: (1) the gender of the leader, (2) the knowledge of the answer to the specific question, (3) the risk attitude of the group member, and (4) the domain type of the question. Although a female leader will increase the willingness of male group members to step forward and answer for the group, the

Table 5: Tobit Estimates Predicting Willingness to Answer: Group with Gender Stereotypes

	Men		Women		Pooled	
	Baseline	Treatment	Baseline	Treatment	Baseline	Treatment
Diff leader	-0.074	1.314*	0.506	-0.022	-0.103	1.175**
	(0.596)	(0.680)	(0.536)	(0.385)	(0.533)	(0.540)
Part A score	0.240	0.158	0.281***	0.060	0.256***	0.096*
	(0.168)	(0.110)	(0.085)	(0.051)	(0.083)	(0.057)
Ans *QiCorr*	1.909***	2.081***	1.948***	2.156***	1.942***	2.125***
	(0.258)	(0.311)	(0.327)	(0.237)	(0.216)	(0.186)
Maleness	0.377	0.643***	-0.729***	-0.480**	0.356	0.646***
	(0.361)	(0.201)	(0.246)	(0.219)	(0.326)	(0.199)
Female					-0.183	0.409
					(0.516)	(0.457)
Diff leader					0.642	-1.245*
× Female					(0.765)	(0.680)
Female					-1.070***	-1.155***
× Maleness					(0.404)	(0.298)
Constant	3.485	-1.061	-0.790	-0.301	-0.484	-0.719
	(3.760)	(3.126)	(2.057)	(3.246)	(1.550)	(2.381)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	420	720	600	960	1,020	1,680

Notes: Control variables include age, session size, risk preferences, and overconfidence. The sum of the coefficients on "Diff leader" and "Diff leader × Female" equals zero (Wald test, $p = 0.852$). For women, coefficients on "Part A score" differ between baseline and treatment ($p = 0.013$, Fisher–Pitman permutation test). Appendix Table 13 reports overconfidence measures. Romano–Wolf adjusted p-values are shown in Appendix Table 20. Similar results from ordered Probit regressions are presented in the Appendix, Table 19. Standard errors clustered at the individual level are in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

effect on group performance needs to be examined.

4.3 Group performance

We start by comparing the performance (i.e., number of questions answered correctly) of female led groups with that of male led groups. Table 6 shows that there is no difference between the male and female led groups at conventional levels. Although, in the only clearly female-typed domain, Art (see Figure 2), female led groups perform marginally better.

Table 6: Observed Group Performance

	Baseline	Treatment		p-value
		Male-led	Female-led	Male-led vs. Female-led
Art	1.88	1.61	2.40	0.06
Geography	2.59	3.06	2.90	0.83
Sports	2.53	3.00	2.70	0.59
Entertainment	3.12	3.39	3.40	1.00
Environmental Science	1.88	1.89	1.70	0.81
History	1.35	1.50	1.60	0.95
Overall	13.35	14.44	14.70	0.88

Notes: Entries report the number of correct answers in each category (maximum of five). The last column reports p-values from Fisher–Pitman permutation tests comparing male-led and female-led groups. Comparisons between the baseline condition and the male-led or female-led treatment groups yield no statistically significant differences.

We now turn to examine how a leader chooses between the group members. In related experiments (Coffman, 2014 and Chen and Houser, 2019), the group member who had a higher willingness would be chosen automatically to answer for the group. In our experiment, it was the leader who made this decision after seeing the willingness to answer for the group together with information on where the group member was sitting (which could then identify their gender in *Treatment*). It is natural to ask how a leader decides who to choose. Recall that there were 30 questions in Part B. For each question the leader needed to choose one of the two group members to answer for the group. We find partial support for Hypothesis 2.

Observation 5 (Leaders' selection). *Leaders are more likely to select the group member who states a higher willingness to contribute. When gender is revealed,*

female leaders are more likely to select female group members.

Support Table 7 reports the average number of times a group member who stated a higher willingness to answer of the two was chosen. On average, the leaders chose the group member who stated a higher willingness to contribute in about 28 of 30 questions in both conditions.[7]

Table 7: Average Number of Answers with Higher Willingness Chosen by Leaders

	N	Mean	Std. dev.
Baseline	17	28.41	1.66
Treatment	28	27.54	2.74

Table 8 reports the results of linear probability models on leaders' choices. The dependent variable is coded as follows: it is 1 if group member 2 was selected, 0 otherwise. "Diff in willingness" is the difference between the willingness of group member 2 and group member 1 (willingness of group member 2 minus willingness of group member 1 at question i). "Diff in ability" is the difference between the two group members' scores in Part A by category. For example, if the leader is going to pick one of the two submitted answers for question i in Art, the "Diff in ability" refers to the difference in their scores in Art in Part A. "Leader Female" and "Member 2 Female" are dummy variables (0 = Male, 1 = Female), indicating the leader's gender and group member 2's gender, respectively. "Maleness" is the average maleness perceived by the leaders. As can be seen in Table 8, the higher the willingness of group member 2 compared to that of group member 1, the more likely the leader chooses group member 2. In addition, and only when the gender is known, a female leader selects a female group member more often.[8,9]

Before we turn to our results on the accuracy of answers conditional on the stated willingness to contribute knowledge, we provide estimates of the prior beliefs held by a set of experts (researchers in experimental and behavioural economics) regarding the accuracy conditional on the stated willingness to contribute knowledge.

[7] In case of a tie, it counted as if the highest willingness was chosen.

[8] We also examine whether the leader might just pick whoever is listed first when the stated willingness is equal. This happens 57.14% (61.89%) of the time in *Baseline (Treatment)*.

[9] When presenting initial results, a male audience member suggested that surely "really good" male leaders are also more likely to choose more capable women. We run additional Logit regressions separately for male leaders at or above the average and those below the average. Results are given in Table 17 in the Appendix. It turns out that "really good" male leaders are even less likely to choose female group members to answer for the group.

Table 8: Probit regressions on Leaders' Choice

	Baseline	Treatment	Baseline	Treatment
Diff in willingness	0.155***	0.154***	0.154***	0.154***
	(0.009)	(0.011)	(0.009)	(0.011)
Diff in ability	0.074***	0.066***	0.073***	0.065***
	(0.012)	(0.009)	(0.012)	(0.010)
Female	-0.010	0.039	-0.010	0.038
	(0.027)	(0.027)	(0.027)	(0.027)
Member 2 Female	-0.036	0.014	-0.036	0.014
	(0.038)	(0.037)	(0.037)	(0.037)
Female	0.068	0.200**	0.067	0.199**
× Member 2 Female	(0.062)	(0.089)	(0.062)	(0.089)
Maleness			0.012	-0.018
			(0.048)	(0.033)
Member 2 Female			-0.056	-0.033
× Maleness			(0.128)	(0.067)
Constant	4.580***	0.481	4.586***	0.484
	(1.128)	(0.621)	(1.096)	(0.620)
Controls	Yes	Yes	Yes	Yes
Observations	510	840	510	840

Notes: Interaction terms are corrected using Norton et al. (2004). The dependent variable is the leader's choice between group member 1 and group member 2. Control variables include age, the leader's risk preferences, and session size. Marginal effects reported. Logit results are similar and reported in Appendix Table 16. Romano–Wolf adjusted p-values are reported in Table 21. All results remain significant after correcting for multiple hypothesis testing. Standard errors clustered at the individual level are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

DellaVigna et al. (2019) argue that collecting predictions before communicating experimental results may mitigate hindsight and publication bias and help assess the novelty and information value of empirical results. In November 2025, before we shared the results from this study, we asked participants at a experimental economics workshop to predict the results. We set up a Qualtrics survey that was distributed to participants prior to the presentation and the best predictions in each category were incentivised with a prize of £5.[10]

Table 9 shows the average predictions. While DellaVigna and Pope (2018) find that the average forecasts of expert researchers typically get the level of an effect as well as the direction right, we do not observe this. While, on average, women are predicted to be more accurate conditional on stating a willingness of 5, there is no perceived difference in accuracy depending on the gender of the leader.

Table 9: Experts' Average Predicted Accuracy When Stating 5 ($N = 35$)

	Male Leadership	Female Leadership	p-value
Male workers	63.80	62.86	0.72
Female workers	71.11	69.17	0.43
p-value	0.03	0.04	

Notes: Entries report experts' average predicted accuracy. p-values are from Fisher–Pitman permutation tests.

Observation 6 (Accuracy and willingness) *Group members who state a higher willingness to answer get more answers correct.*

Support Table 10 reports the accuracy of group members' submitted answers conditional on their reported willingness by gender. The willingness to answer increases with the accuracy. In *Baseline* as well as in *Treatment*, group members who are most eager to answer for the group (i.e., willingness = 5) are the ones who are most accurate. Interestingly, conditionally on stating 5, women and men are equally correct in *Baseline* ($p = 0.31$, Test of Proportions).[11] However, there are distinct effects emerging when gender is made salient.

Observation 7 (Accuracy of highly willing men). *Conditional on stating the highest willingness to contribute, male group members are less accurate when*

[10]For a description of the experiment and the precise prediction questions see section 6.3 in the Appendix.

[11]Whenever we refer to the Test of Proportions, we conduct clustered tests at the individual level, since each individual makes 30 decisions.

Table 10: Answer Accuracy in Part B Conditional on Stated Willingness, by Gender

	Overall		Treatment	
Willingness	Baseline	Treatment	Male-led	Female-led
Male				
1	23.81%	20.61%	15.09%	30.51%
2	26.76%	33.33%	38.36%	27.12%
3	29.76%	34.35%	44.29%	22.95%
4	30.67%	39.39%	42.50%	37.29%
5	82.35%	68.39%	**76.24%**	**59.78%**
Female				
1	20.93%	24.48%	27.38%	17.81%
2	28.99%	27.98%	23.42%	36.47%
3	34.86%	32.94%	31.62%	38.24%
4	46.38%	50.45%	55.29%	34.62%
5	75.89%	85.64%	**81.82%**	**96.15%**

Notes: Willingness ranges from 1 (weak) to 5 (strong). The table reports the accuracy of male and female group members' submitted answers. "Male-led" ("Female-led") refers to participants assigned to male (female) leaders.

paired with a female leader than when paired with a male leader or when leader gender is unknown.

Support In *Treatment*, the accuracy of men who state a willingness of 5 when facing a female leader is significantly lower than the accuracy of men who face a male leader (60% vs. 76%, $p = 0.02$, Test of Proportions). This cannot be interpreted as in-group favouritism (brotherhood effect) as the conditional accuracy when facing a male leader is not different from the conditional accuracy when not knowing the gender of the leader (76% vs. 82%, $p = 0.34$, Test of Proportions). The conditional accuracy drops significantly when facing a female leader compared to when one does not know the gender of the leader (60% vs. 82%, $p < 0.01$, Test of Proportions). This male reaction is only observed in the presence of a female leader.

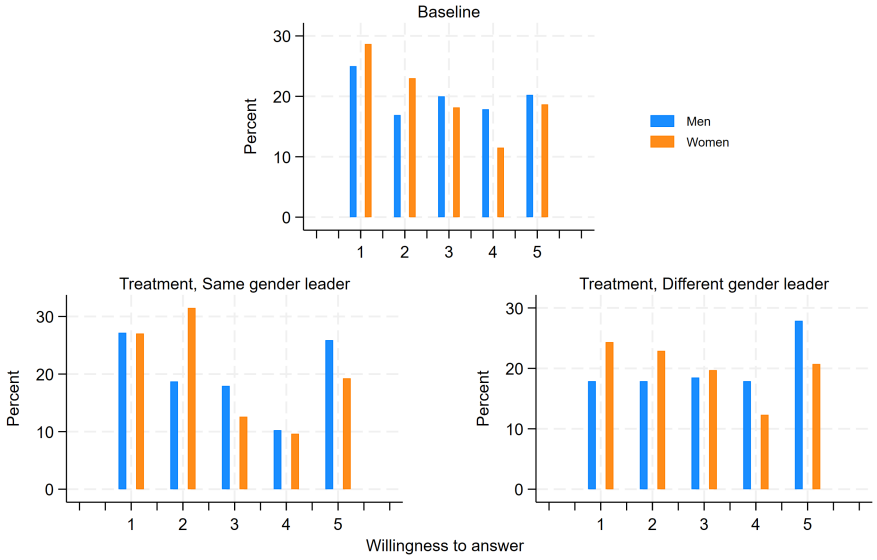
Observation 8 (Accuracy of highly willing women). *Conditional on stating the highest willingness to contribute, female group members are more accurate when paired with a female leader than when paired with a male leader or when leader gender is unknown.*

Support For women who state a willingness of 5 and know they are facing a female leader, their accuracy is 96% which is significantly higher than when they are facing a male leader (82%) ($p = 0.02$, Test of Proportions). Women who face a male leader exhibit no significant differences to those who do not know the gender of the leader (82% vs. 76%, $p = 0.28$, Test of Proportions). What drives the observed sisterhood effect is the significant increase in the conditional accuracy when facing a female leader (96% vs. 76%). Women only state the highest willingness to step forward and contribute their knowledge when they are almost certain they get the answer right.[12] Our experts correctly predicted a general desire for accuracy by women, i.e., women are predicted to get more answers right than men when expressing the highest willingness to contribute their knowledge (Table 9). However, the super correctness when facing a female leader (*Sisterhood effect*), indicative of a desire not to let the female leader down, was not predicted by the experts.

Figure 3 shows the distribution of willingness for group members in *Baseline* as well as *Treatment*. The distribution of the stated willingness of men facing a female leader is significantly different from that of men facing a male leader in *Treatment* ($p = 0.05$, Kolmogorov–Smirnov test). Similar results are found among women

[12]Practitioner accounts sometimes claim that men are willing to put themselves forward with lower perceived readiness than women, often summarized as a "60% versus 100%" threshold. We cite this only to note a superficial similarity in magnitudes. Our contribution is to document accuracy conditional on stated willingness in a controlled laboratory setting, not to validate practitioner narratives.

Figure 3: Distribution of willingness to contribute knowledge



under female leadership ($p = 0.014$, Kolmogorov–Smirnov test). The distributions of the stated willingness are different between those facing male and female leaders ($p = 0.01$, Kolmogorov–Smirnov test). To be more specific, men are more likely to state a (moderate) strong willingness to contribute when matched with a different gender leader (i.e., are shifting mass to the right of the distribution); women are less likely to submit a (moderate) weak willingness to contribute when their leaders are male (i.e., are shifting mass to the middle of the distribution). Looking at the distributions, it seems that if the stated willingness can be assumed of as a signal, both men and women tend to send a more bipolar ("precise") signal to a leader of the same gender, it is more "fuzzy" when it is sent to a different gender leader.

In addition to analysing accuracy conditional on stated willingness, we examine whether leader gender affects overall answer accuracy in Part B. Appendix Table 18 reports OLS regressions of individual performance on Part B questions. Other than individual performance in Part A, no covariates are statistically significant. In particular, neither the treatment indicator nor its interaction with a different-gender leader predicts overall accuracy.

This indicates that leader gender does not directly affect cognitive performance or effort. Instead, leader gender affects the accuracy threshold at which individuals are willing to translate private knowledge into expressed willingness to contribute,

without changing average ability.

4.3.1 *A simple counterfactual selection rule*

Our results suggest that the informational content of stated willingness depends on leader gender: conditional on maximum willingness, women's answers are especially accurate under female leadership, whereas men's answers are less accurate under female leadership. This raises a practical question: could a simple rule-of-thumb improve performance in environments where leaders face noisy signals of contribution quality?

To illustrate, we evaluate a parsimonious counterfactual rule that uses only information available to the leader in our setting. In the simulations, the leader is forced to select the female group member whenever she states the maximum willingness (5). If no female member states 5, the leader selects the member with the higher stated willingness; ties are broken as in the data. We implement this rule using the experimentally observed answers, willingness reports, and leader–member gender in *Treatment*, and compute group scores under repeated re-draws. We run 10,000 repetitions.

Table 11: Comparison of Simulated and Observed Experimental Group Performance

	Women First	Treatment	Baseline
Male leaders	14.80	14.44	13.44
Female leaders	13.67	14.70	13.25

Notes: Simulated results are based on 10,000 repetitions.

Table 11 reports observed and simulated group performance. The counterfactual rule increases performance in male-led groups relative to the experimental benchmark (Fisher–Pitman permutation test, $p < 0.01$). For female-led groups, the rule yields no improvement, consistent with female leaders already selecting women more often when gender is revealed.

4.3.2 *Interpretation*

When leader gender is revealed, Observations 7 and 8 reveal a pronounced asymmetry in the accuracy threshold at which men and women are willing to state the

maximum willingness. Men's conditional accuracy drops sharply under female leadership, while women's conditional accuracy rises.

This pattern is difficult to reconcile with changes in underlying ability or effort. Instead, it suggests that leader gender alters how individuals map strong private signals into expressed willingness to contribute. Women appear to apply a stricter accuracy threshold under female leadership, whereas men apply a weaker threshold.

One interpretation is that women's behaviour is consistent with in-group discipline under female leadership, while men's behaviour reflects a negative response to female leadership relative to the benchmark cases of male leadership and leader gender being unknown. Importantly, this asymmetry does not reflect changes in overall cognitive performance: we find no effect of leader gender on unconditional answer accuracy (Appendix Table 18).

A key design feature is that gender is made salient via a seating-based cue. In principle, behaviour in *Treatment* could therefore reflect a generic "square versus triangle" categorization rather than gender per se. The comparison with *Baseline* provides an important check: behaviour toward male leaders in *Treatment* is statistically indistinguishable from behaviour when leader gender is not revealed, whereas behaviour changes sharply when the leader is female. This pattern is difficult to reconcile with a symmetric minimal-group mechanism and instead points to a response triggered specifically by female leadership. Consistent with this interpretation, we find that male group members are *less* willing to state high willingness under female leadership when they know their answer is correct, indicating a disruption in how private information is translated into participation rather than a uniform increase or decrease in effort.[13]

Although fear of failure could contribute to women's behaviour more generally, our setting points to a distinct mechanism. Contributing an incorrect answer affects not only individual payoffs but also group outcomes and the leader's payoff. Women may therefore be especially selective about stating the highest willingness under female leadership, consistent with a desire not to impose costs on an in-group leader. By contrast, the lower conditional accuracy of men under female leadership suggests a relaxation of the accuracy threshold at which a high willingness is expressed.

[13]We examine whether male group members' willingness to contribute depends on leader gender, conditional on knowing the correct answer. Appendix Table 23 reports Tobit regressions interacting answer correctness with leader gender. Male group members are less willing to contribute under female leadership when their answer is correct, while no such effect is observed among female group members. This pattern suggests that female leadership alters how men map strong private signals into expressed willingness, rather than uniformly increasing or decreasing participation.

Finally, the absence of meaningful differences between facing a male leader and not knowing leader gender is consistent with male leadership functioning as the salient default in this setting. When leader gender is revealed to be female, both genders adjust behaviour, but in opposite directions: men lower, and women raise, the accuracy threshold at which they are willing to state the highest willingness to contribute.

5 Conclusion

In organizations where teamwork is central, leader gender may shape not only who speaks up, but also the informational quality of what is offered and ultimately selected. This paper studies how leader gender affects both the *supply* of contributions within a group and the *selection* of contributions for collective decisions. We conduct a controlled laboratory experiment in which participants are randomly assigned to groups of three consisting of one merit-based leader and two group members. Group members answer questions and report their willingness—on a scale from one to five—to have their answer implemented as the group response, after which the leader selects one answer per question. In a baseline condition, gender is not revealed; in a treatment condition, leader and member gender are made salient. This design allows us to separate how leader gender affects willingness to step forward from how it affects whose contributions are chosen.

Our first main finding concerns the supply of ideas. When leader gender is revealed, men become more willing to contribute under female leadership than under male leadership or when leader gender is unknown. However, this increase in participation is accompanied by a decline in the accuracy threshold at which men are willing to state the highest willingness. Women's willingness to contribute, by contrast, does not increase under female leadership, but conditional on stating maximum willingness, women's answers are substantially more accurate under female leaders than under male leaders or when leader gender is unknown. We refer to this asymmetric pattern—higher conditional accuracy among highly willing women and lower conditional accuracy among highly willing men under female leadership—as a *sisterhood effect*. We do not observe a corresponding improvement in men's conditional accuracy under male leadership relative to the no-gender benchmark.

Importantly, these effects do not reflect changes in underlying ability or overall performance. Leader gender does not affect average accuracy or effort; instead, it shifts the accuracy threshold at which individuals are willing to translate private

knowledge into expressed willingness to contribute. Individuals perform similarly across conditions, but apply different standards when deciding whether to put their answer forward with maximal willingness.

This asymmetry contrasts with the symmetric in-group responses predicted by social identity theories of leadership, and instead points to a role for leadership norms in which male leadership functions as a default while female leadership activates identity-relevant adjustments. One interpretation of this asymmetry is that making female leadership salient activates identity-relevant cues that affect men and women differently, rather than generating symmetric in-group favoritism. Experimental evidence from social psychology points in this direction. Yuki and Yokota (2009) show that subtle outgroup threat primes increase intergroup discrimination among men but not among women, even when the task itself is unrelated to group competition. Their findings suggest that men's behaviour is more responsive to contextual cues that render social identity or group boundaries salient, whereas women's responses are comparatively stable. In our setting, revealing female leadership may similarly function as an identity-relevant cue that shifts men's willingness standards—lowering the accuracy threshold at which confidence is expressed—without generating a corresponding adjustment among women. This interpretation is consistent with the asymmetric "sisterhood effect" we observe and contrasts with models that predict symmetric in-group responses to leader gender.

Our second main finding concerns the selection of ideas. Leaders place substantial weight on stated willingness when choosing which answer represents the group. In addition, when gender is revealed, female leaders are more likely to select female group members, whereas male leaders show no systematic gender-based selection. Importantly, the lower conditional accuracy of highly willing men under female leadership does not translate into lower group performance, consistent with female leaders' selection behaviour filtering out lower-quality contributions. Motivated by this pattern, we simulate a parsimonious counterfactual rule—select the female member whenever she states willingness equal to five; otherwise select the higher willingness—using only information available to leaders in our setting. This rule improves performance in male-led groups but yields no improvement in female-led groups, consistent with female leaders already selecting in a way that effectively leverages the informational content of willingness.

A unifying feature of these findings is that leader gender reshapes the informational environment not by changing what group members know, but by changing how knowledge is expressed and filtered. Leader gender affects both the mapping from private signals into stated willingness and the mapping from stated willingness

into selection. These two margins—expression and selection—jointly determine how information is aggregated in groups.

Taken together, our results highlight two distinct ways in which leader gender can matter for group decision-making. Female leadership can expand the pool of contributions by increasing men's willingness to step forward, while simultaneously preserving decision quality through selection behaviour that limits the influence of low-quality, high-confidence submissions. More broadly, leader gender affects collective performance through systematic changes in both the mapping from private knowledge to expressed willingness and the mapping from willingness to selection.

Our findings also speak to the literature on backlash against female leaders. Prior work documents that female leaders often receive less support, are evaluated more harshly, and face resistance from subordinates even when performance is held constant (e.g., Abel, 2024; Ayalew et al., 2021; Boring, 2017; Grossman et al., 2019). We document a distinct mechanism through which such backlash can operate: not through reduced participation, but through a deterioration in the informational quality of contributions by male subordinates. When led by a woman, men become more willing to contribute, yet do so at a substantially lower accuracy threshold, injecting noise into the information aggregation process. This response is asymmetric and does not arise under male leadership. Female leaders partially offset this mechanism through selective filtering, helping explain why group performance does not deteriorate despite increased noise.

Our study has limitations. The use of student participants and quiz-based tasks may limit external validity, and future work should examine whether similar patterns arise in professional settings and in environments where contribution quality is less verifiable ex post. Leadership in our experiment is merit-based and interaction is short-lived; how feedback, repeated interaction, and alternative pathways to leadership moderate these effects remains an important direction for future research. Despite these limitations, the experiment provides evidence that leader gender can systematically affect participation, contribution quality conditional on expressed willingness, and the aggregation of information in groups.

References

Abel, M. (2024). Do workers discriminate against female bosses? *Journal of Human Resources*, *59*(2), 470–501.

Ahern, K. R., & Dittmar, A. K. (2012). The changing of the boards: The impact on firm valuation of mandated female board representation. *The Quarterly Journal of Economics*, *127*(1), 137–197.

Ayalew, S., Manian, S., & Sheth, K. (2021). Discrimination from below: Experimental evidence from Ethiopia. *Journal of Development Economics*, *151*, 102653.

Beaman, L., Duflo, E., Pande, R., & Topalova, P. (2012). Female leadership raises aspirations and educational attainment for girls: A policy experiment in india. *Science*, *335*(6068), 582–586.

Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, *131*(4), 1753–1794.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, *145*, 27–41.

Carter, D. A., Simkins, B. J., & Simpson, W. G. (2003). Corporate governance, board diversity, and firm value. *Financial Review*, *38*(1), 33–53.

Chakraborty, P., & Serra, D. (2024). Gender and leadership in organisations: The threat of backlash. *The Economic Journal*, *134*(660), 1401–1430.

Charness, G., & Chen, Y. (2020). Social identity, group behavior, and teams. *Annual Review of Economics*, *12*(1), 691–713.

Chen, J., & Houser, D. (2019). When are women willing to lead? The effect of team gender composition and gendered tasks. *The Leadership Quarterly*, *30*(6), 101340.

Clarke, D., Romano, J. P., & Wolf, M. (2020). The Romano–Wolf multiple-hypothesis correction in stata. *The Stata Journal*, *20*(4), 812–843.

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, *129*(4), 1625–1660.

DellaVigna, S., & Pope, D. (2018). Predicting experimental results: Who knows what? *Journal of Political Economy*, *126(6)*, 2410–2456.

DellaVigna, S., Pope, D., & Vivalt, E. (2019). Predict science to improve science. *Science*, *366*, 428–429.

Evans, R. B., Prado, M. P., Rizzo, A. E., & Zambrana, R. (2024). Identity, diversity, and team performance: Evidence from US mutual funds. *Management Science*.

Faccio, M., Marchica, M.-T., & Mura, R. (2016). CEO gender, corporate risk-taking, and the efficiency of capital allocation. *Journal of Corporate Finance*, *39*, 193–209.

Fischbacher, U. (2007). Z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178.

Gangadharan, L., Jain, T., Maitra, P., & Vecci, J. (2016). Social identity and governance: The behavioral response to female leaders. *European Economic Review*, *90*, 302–325.

Grossman, P. J., Eckel, C., Komai, M., & Zhan, W. (2019). It pays to be a man: Rewards for leaders in a coordination game. *Journal of Economic Behavior & Organization*, *161*, 197–215.

Grossman, P. J., Komai, M., & Jensen, J. E. (2015). Leadership and gender in groups: An experiment. *Canadian Journal of Economics/Revue canadienne d'économique*, *48*(1), 368–388.

Hamilton, B. H., Nickerson, J. A., & Owan, H. (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *Journal of Political Economy*, *111*(3), 465–497.

Haslam, S. A., Reicher, S. D., & Platow, M. J. (2015). Leadership: Theory and practice. *APA handbook of personality and social psychology, Volume 2: Group processes.*, 67–94.

Hogg, M. A., & Van Knippenberg, D. (2003). Social identity and leadership processes in groups. *Advances in Experimental Social Psychology*, *35*, 1–52.

Hogg, M. A., Van Knippenberg, D., & Rast III, D. E. (2012). The social identity theory of leadership: Theoretical origins, research findings, and conceptual developments. *European Review of Social Psychology*, *23*(1), 258–304.

Kanter, R. M. (1977). Some effects of proportions on group life: Skewed sex ratios and responses to token women. *American Journal of Sociology*, *82*(5), 965–990.

Karni, E. (2009). A mechanism for eliciting probabilities. *Econometrica*, *77*(2), 603–606.

Karpowitz, C. F., O'Connell, S. D., Preece, J., & Stoddard, O. (2024). Strength in numbers? Gender composition, leadership, and women's influence in teams. *Journal of Political Economy*, *132*(9), 3077–3114.

Matsa, D. A., & Miller, A. R. (2011). Chipping away at the glass ceiling: Gender spillovers in corporate leadership. *American Economic Review*, *101*(3), 635–39.

Matsa, D. A., & Miller, A. R. (2013). A female style in corporate leadership? Evidence from quotas. *American Economic Journal: Applied Economics*, *5*(3), 136–69.

Möbius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2022). Managing self-confidence: Theory and experimental evidence. *Management Science*, *68*(11), 7793–7817.

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, *122(3)*, 1067–1101.

Norton, E. C., Wang, H., & Ai, C. (2004). Computing interaction effects and standard errors in logit and probit models. *The Stata Journal*, *4*(2), 154–167.

Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, *73*(4), 1237–1282.

Romano, J. P., & Wolf, M. (2016). Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters*, *113*, 38–40.

Tajfel, H., Turner, J., Austin, W. G., & Worchel, S. (2001). An integrative theory of intergroup conflict. *Intergroup relations: Essential readings*, 94–109.

Yuki, M., & Yokota, K. (2009). The primal warrior: Outgroup threat priming enhances intergroup discrimination in men but not women. *Journal of Experimental Social Psychology*, *45*(1), 271–274. https://doi.org/https://doi.org/10.1016/j.jesp.2008.08.018

Zhang, L. (2020). An institutional approach to gender diversity and firm performance. *Organization Science*, *31*(2), 439–457.

6 Appendix

6.1 Additional results

Table 12: Performance of All Participants in Part A

Category	Men	Women	p-value (M = W)
Art	2.44	2.33	0.61
Geography	2.83	2.27	0.01
Sports	2.89	2.53	0.03
Entertainment	1.15	1.29	0.49
Environmental Science	3.94	3.61	0.09
History	1.95	1.66	0.15
N	65	70	

Notes: p-values are from Fisher–Pitman permutation tests.

Table 13: Group Members' Expected and Actual Scores in Part A

	Men		Women	
Category	Expected	Actual	Expected	Actual
Art	2.69***	1.89	2.73***	2.10
Geography	3.04***	2.32	2.89***	2.06
Sports	2.74	2.58	2.17	2.37
Entertainment	1.99***	1.13	1.89***	1.13
Environmental Science	3.41	3.76	3.38	3.44
History	2.51***	1.55	2.15***	1.38
Overall	16.38***	13.23	15.21***	12.48

Notes: Expected scores are computed from elicited beliefs in Part A. Fisher–Pitman permutation tests are used to compare expected and actual scores separately for men and women. If a group member's expected overall score exceeds the actual overall score, the individual is classified as overconfident. Men expect to score higher than women in Sports ($p < 0.01$) and History ($p < 0.05$). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 14: Group Performance by Category

Category	Baseline	Treatment	p-value
Art	1.88	1.89	1.00
Geography	2.59	3.00	0.22
Sports	2.53	2.89	0.39
Entertainment	3.12	3.39	0.50
Environmental Science	1.88	1.82	0.96
History	1.35	1.54	0.70
Overall	13.35	14.54	0.23
Observations	17	28	

Notes: Performance ranges from 0 to 5 at the category level. p-values are from Fisher–Pitman permutation tests.

Table 15: Guess of group performance by group members

Category	Baseline	Treatment	p-value
Art	2.68	2.68	1.00
Geography	3.09	3.18	0.72
Sports	3	3.05	0.86
Entertainment	3.38	3.30	0.80
Environmental Science	2.74	2.85	0.65
History	2.94	2.70	0.18
Overall	17.82	17.77	0.96

Notes: Group members are asked to guess their group's scores in each category. p-values are from Fisher–Pitman permutation tests.

Table 16: Logit Regressions on Leaders' Choice

	Baseline	Treatment	Baseline	Treatment
Diff in willingness	1.689***	1.121***	1.691***	1.119***
	(0.259)	(0.180)	(0.262)	(0.181)
Diff in ability	0.763***	0.456***	0.756***	0.453***
	(0.126)	(0.071)	(0.129)	(0.073)
Female	-0.508	-0.607	-0.515	-0.606
	(0.569)	(0.504)	(0.545)	(0.502)
Member 2 Female	-0.631	-0.373	-0.609	-0.360
	(0.618)	(0.323)	(0.616)	(0.333)
Female	0.680	1.164*	0.680	1.152*
× Member 2 Female	(0.632)	(0.620)	(0.609)	(0.617)
Maleness			0.505	-0.030
			(1.331)	(0.357)
Maleness			-0.639	-0.184
× Member 2 Female			(1.412)	(0.480)
Constant	8.074***	0.598	8.073***	0.588
	(2.050)	(1.165)	(2.019)	(1.162)
Controls	Yes	Yes	Yes	Yes
Observations	510	840	510	840

Notes: Control variables include age, the leader's risk preferences, and session size. Standard errors clustered at the individual level are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 17: Logit Regressions on Male Leaders' Choice by Ability

	High	Low
Diff in willingness	1.948***	1.420***
	(0.582)	(0.515)
Diff in ability	0.473**	0.585***
	(0.192)	(0.200)
Member 2 Female	-1.097***	1.009***
	(0.425)	(0.248)
Maleness	-1.087	0.488*
	(1.110)	(0.278)
Maleness	0.878	-0.632
× Member 2 Female	(1.201)	(0.653)
Constant	-1.695	2.098
	(4.124)	(2.668)
Controls	Yes	Yes
Observations	330	210

Notes: Male leaders are categorized into High and Low types based on performance in Part A. Control variables include age, session size, and the leader's risk preferences. Standard errors clustered at the individual level are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 18: OLS Predicting Group Members' performance on Part B questions

	(1)	(2)	(3)
Treatment	0.482	0.380	0.237
	(1.009)	(1.001)	(0.881)
Diffleader	-0.679	-0.726	-0.859
	(1.051)	(1.041)	(0.868)
Treatment × Diffleader	0.157	0.143	-0.028
	(1.340)	(1.327)	(1.136)
Part A scores	0.484***	0.505***	0.268**
	(0.135)	(0.134)	(0.128)
Risk		-0.324	-0.237
		(0.197)	(0.169)
Constant	5.985**	7.544***	12.41**
Controls	No	No	Yes
Observations	90	90	90

Notes: The dependent variable is the number of correct answers submitted by group members in Part B. Control variables include age, session size, and ethnicity. Standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 19: Ordered Probit Regressions Predicting Willingness to Contribute

	Men		Women		Pooled	
	Baseline	Treatment	Baseline	Treatment	Baseline	Treatment
Diff leader	-0.040	0.539**	0.206	0.004	-0.052	0.526**
	(0.268)	(0.272)	(0.232)	(0.185)	(0.240)	(0.236)
Part A score	0.107	0.064	0.128***	0.041*	0.117***	0.048**
	(0.075)	(0.045)	(0.038)	(0.022)	(0.037)	(0.024)
Ans *QiCorr*	0.862***	0.833***	0.875***	0.995***	0.867***	0.923***
	(0.122)	(0.107)	(0.107)	(0.097)	(0.080)	(0.071)
Maleness	0.179	0.272***	-0.315***	-0.226**	0.171	0.296***
	(0.167)	(0.081)	(0.105)	(0.107)	(0.149)	(0.089)
Female					-0.091	0.167
					(0.234)	(0.204)
Diff leader					0.275	-0.540*
× Female					(0.338)	(0.301)
Female					-0.480***	-0.521***
× Maleness					(0.180)	(0.136)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	420	720	600	960	1,020	1,680

Notes: Control variables include age, session size, risk preferences, and overconfidence. Standard errors clustered at the individual level are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 20: Adjusted p-Values for Tables 4 and 5

	Men		Women		Pooled	
	Baseline	Treatment	Baseline	Treatment	Baseline	Treatment
Panel A. Table 4						
Coefficient (Diff leader)	-0.219	1.176	0.230	-0.014	-0.216	1.072
p value	0.6784	0.0447	0.6893	0.9725	0.6872	0.0318
Romano-Wolf p value	0.6384	0.0010	0.6454	0.9291	0.6454	0.0010
Coefficient (Diff leader × Female)					0.476	-1.127
p value					0.5414	0.0963
Romano-Wolf p value					0.3497	0.0020
Panel B. Table 5						
Coefficient (Diff leader)	-0.074	1.314	0.506	-0.022	-0.103	1.175
p value	0.9009	0.0538	0.3448	0.9552	0.8468	0.0297
Romano-Wolf p value	0.9520	0.0020	0.1089	0.9520	0.9181	0.0010
Coefficient (Diff leader × Female)					0.642	-1.245
p value					0.4021	0.0674
Romano-Wolf p value					0.1598	0.0030

Notes: Romano–Wolf adjusted p-values are based on 1,000 bootstrap replications.

Table 21: Adjusted p-Values for Table 8

	Baseline	Treatment	Baseline	Treatment
Coefficient (Female × Member 2 Female)	0.385	0.803	0.378	0.796
p value	0.2592	0.0205	0.2611	0.0208
Romano-Wolf p value	0.2557	0.0310	0.2557	0.0310

Notes: Romano–Wolf adjusted p-values are based on 1,000 bootstrap replications. The coefficients come from probit regressions without the correction (Norton et al., 2004).

Table 22: Tobit Regressions Predicting Willingness Interacting with Maleness

	Men		Women		Pooled	
	Control	Treatment	Control	Treatment	Control	Treatment
Diff leader	-0.072	1.312*	0.490	0.004	-0.100	1.174**
	(0.602)	(0.693)	(0.511)	(0.382)	(0.537)	(0.554)
Maleness	0.392	0.634**	-0.864**	-0.230	0.370	0.635**
	(0.643)	(0.317)	(0.356)	(0.474)	(0.613)	(0.306)
Part A score	0.240	0.158	0.286***	0.062	0.258***	0.097*
	(0.168)	(0.110)	(0.086)	(0.052)	(0.083)	(0.057)
Ans *QiCorr*	1.908***	2.081***	1.950***	2.164***	1.944***	2.129***
	(0.255)	(0.312)	(0.330)	(0.236)	(0.217)	(0.186)
Diff leader	-0.030	0.019	0.241	-0.348	-0.034	0.022
× Maleness	(0.650)	(0.425)	(0.496)	(0.530)	(0.640)	(0.407)
Female					-0.172	0.390
					(0.511)	(0.458)
Diff leader					0.622	-1.218*
× Female					(0.753)	(0.689)
Female					-1.210*	-0.897
× Maleness					(0.698)	(0.581)
Diff leader					0.258	-0.364
× Female × Maleness					(0.800)	(0.685)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
420	720	600	960	1,020	1,680	

Notes: Control variables include age, session size, risk preferences, and overconfidence. Standard errors clustered at the individual level are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 23: Tobit predicting willingness with accuracy at question level

	Men		Women	
	Control	Treatment	Control	Treatment
Diff leader	-0.025	1.830**	0.699	0.231
	(0.642)	(0.770)	(0.594)	(0.465)
Ans *QiCorr*	1.973***	2.662***	2.227***	2.593***
	(0.404)	(0.447)	(0.488)	(0.494)
Diff leader	-0.136	-1.301**	-0.509	-0.610
× Ans *QiCorr*	(0.488)	(0.538)	(0.528)	(0.516)
Part A score	0.238	0.154	0.276***	0.063
	(0.170)	(0.106)	(0.084)	(0.052)
Maleness	0.377	0.607***	-0.734***	-0.460**
	(0.361)	(0.187)	(0.246)	(0.216)
Controls	Yes	Yes	Yes	Yes
Observations	420	720	600	960

Notes: Control variables include age, session size, risk preferences, and overconfidence. Standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

6.2 Experimental instructions

Welcome to the experiment.

Please do not use your electronic devices/phones brought with you during the experiment. Using the internet during the experiment is not allowed. Anyone who violates these rules will be dismissed without payment. If you have questions, please raise your hand.

Please read and review the consent forms on your desks. If you wish to participate, please sign the consent forms.

There are four parts containing five tasks in this experiment. In each part, you have the opportunity to get points. Each point will earn you 50 pence. One of these five tasks will be randomly chosen to determine your payment. At the end of the experiment, you will be informed of your payment and the task which determines your payment. The showup fee is included in the payment.

Part A:

There are two tasks in this part.

Task 1 In this task, you need to answer questions in six categories: Sports, History, Geography, Environmental Science, Art and Entertainment. There are five multiple-choice questions in each category.

You will receive 1 point for every question you answer correctly. There is no penalty for incorrect answers.

Task 2: In this task, you are asked to estimate the probability of your answer being correct in each question when you are answering the question. We employ 100 different robots to help you. Each Robot has an accuracy corresponding to an integer between 1 and 100. That is, Robot 1 is accurate 1% of the time, Robot 2 is accurate 2% of the time, Robot 3 is accurate 3% of the time, ...all the way up to Robot 100 which is accurate 100% of the time. We will randomly pick one of the 100 different Robots and match it with you. For each question, if your estimated probability is greater than the Robot's accuracy level, your answer will be submitted. If the Robot's accuracy level is greater than your estimated probability, the Robot will answer the question for you.

For example, suppose that you think you are 60% confident of your answer to the question and the Robot you are matched with is Robot 34; then your answer for the question will be submitted since you think you are more likely to give the

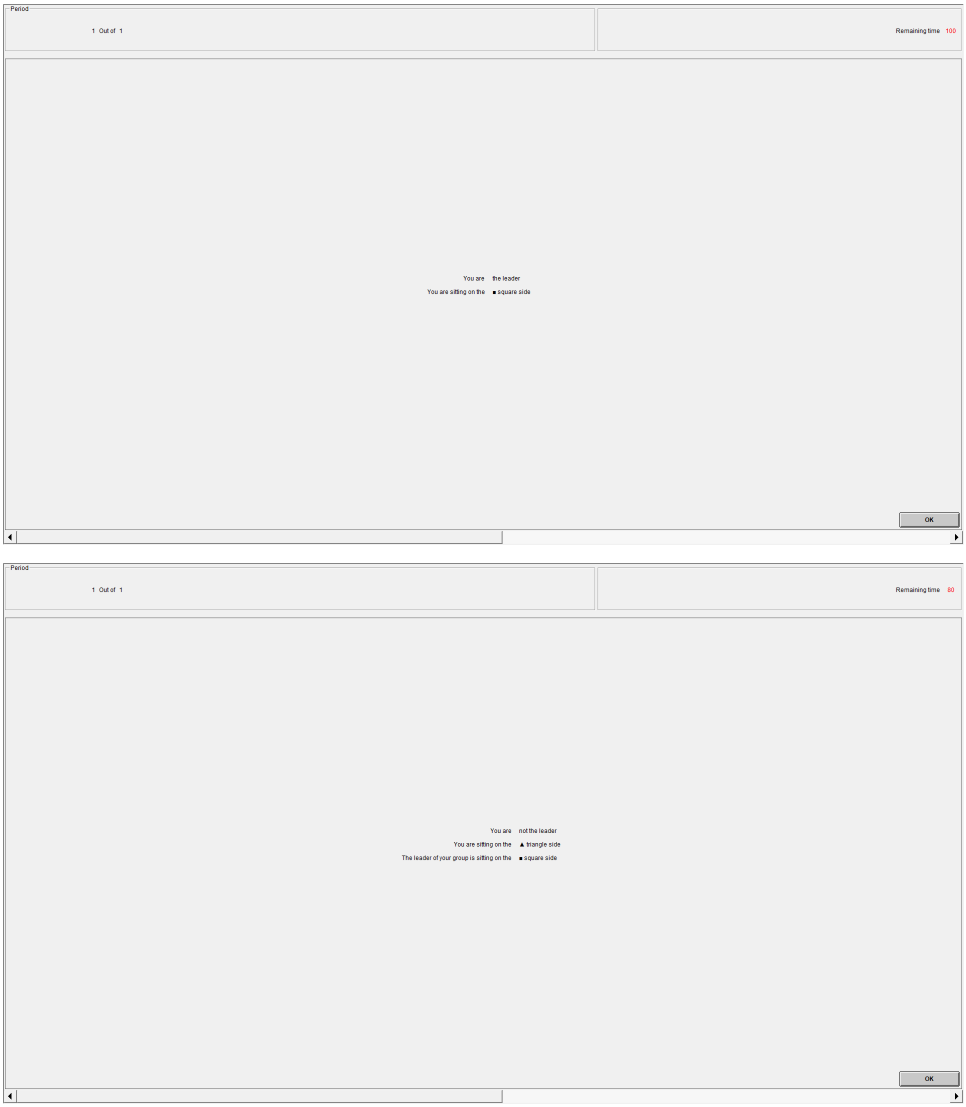
correct answer for the question than the Robot. But if the Robot you are matched with is Robot 97, then we will use the Robot's answer since the Robot is more likely to give the correct answer for the question. You will not be informed which Robot is matched with you.

If this part is selected to determine your payment. Each correct answer, no matter if it comes from you or the Robot, will earn you one point. There is no penalty for incorrect answers.

If you have any questions, please raise your hand.

Part B:

In this part, you will be randomly assigned into a group of three. There are two group members and one group leader in the group. You might be a group member or a leader. Everyone in the group is sitting in this lab. The experimenter will randomly assign the participants who are the best 1/3 in Part A, Task 1, into each group as the leader.



In a group, the two group members will need to submit their answers and their willingness to answer for the group in each question. The leader will need to choose one of the two group members to answer for the group for each question. For each question, the group member chosen by the leader will receive 0.1 points. If the answer to the question is correct, everyone in the group will receive one point. Otherwise, everyone in the group will receive nothing. Let's see how this works in detail.

Firstly, in a group, the two group members are going to answer new questions in six categories. There are five questions in each category. Questions in this part are different to those in Part A. The group members will be informed of the side

their leader is sitting on.

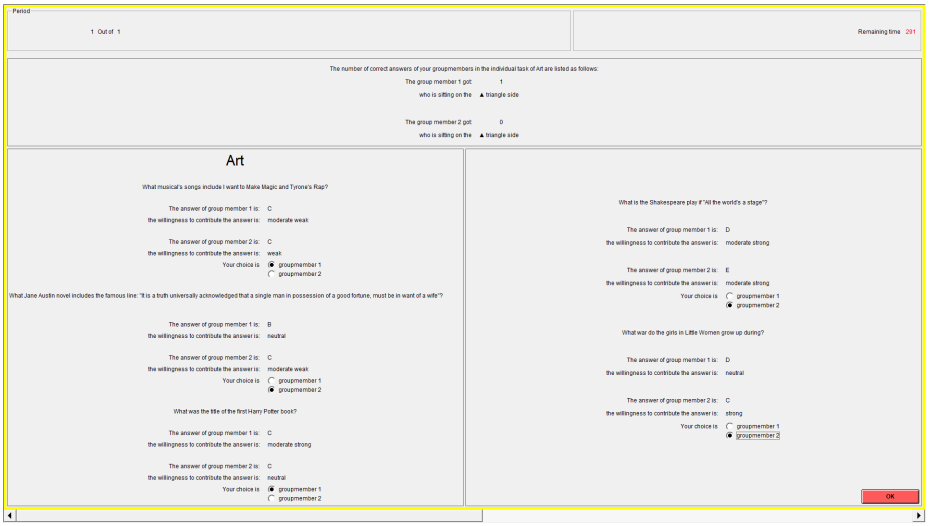
For each question, group members will need to answer the question and submit their willingness to answer for the group by choosing a number from 1 to 5.

1 – the weak willingness to answer for the group; 2 – the moderate weak willingness to answer for the group; 3 – to be neutral; 4 – the moderate strong willingness to answer for the group; 5 – the strong willingness to answer for the group;

After group members finish all the questions, their leader needs to decide who will answer for their group in each question.

Some information will be available to the leader when making decisions:

1. The performance of your group members in the first task 2. The submitted answers from the two group members 3. Group members' willingness to answer for the group 4. On which side their group members are sitting on

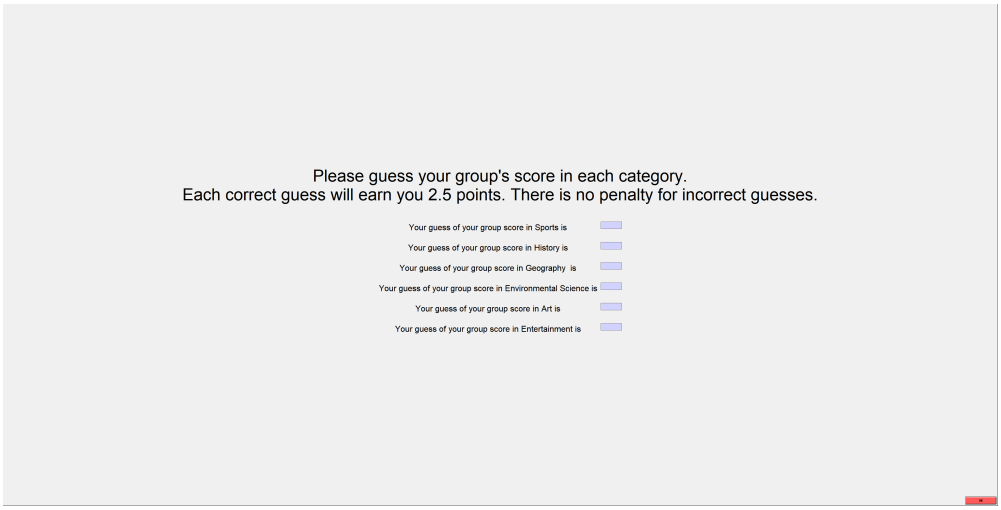


Alternative answers for each question will not be presented to the leader. In each question, the group member chosen by the leader will receive 0.1 points, no matter whether the answer is correct or not. Besides, if that group member answers the question correctly, everyone in the group will receive one point. In addition, the leader will receive 5 points as a reward. There is no penalty for incorrect answers.

If you have any questions, please raise your hand. If you are ready, you may press the button.

Part C:

In this part, you need to guess how many questions your group answered correctly in Part B by categories. Reminding that in Part B, there are 30 questions in six categories, 5 in each. You will receive 2.5 points for each correct guess. There is no penalty for incorrect guesses. If you have any questions, please raise your hand. If you are ready, you may start to guess and submit your guess.



Part D:

In this part, you need to choose between two different options, Option A and Option B. If you choose Option A, you will get 0 points for sure. If you choose Option B, the number of points you receive will depend on a lottery. The lottery works like this: The computer is going to draw a random number between 1 and 100. If the number is less than or equal to the threshold X, you will win 1 point. If the randomly-drawn number is greater than the threshold X, you lose 1/4 of a point.

For example, if the threshold X is 40 and you choose Option B, the lottery, and the randomly drawn number is 65, you would lose 1/4 of a point because you chose the lottery, and 65 is greater than 40. If the randomly drawn number is 35, you will receive 1 point.

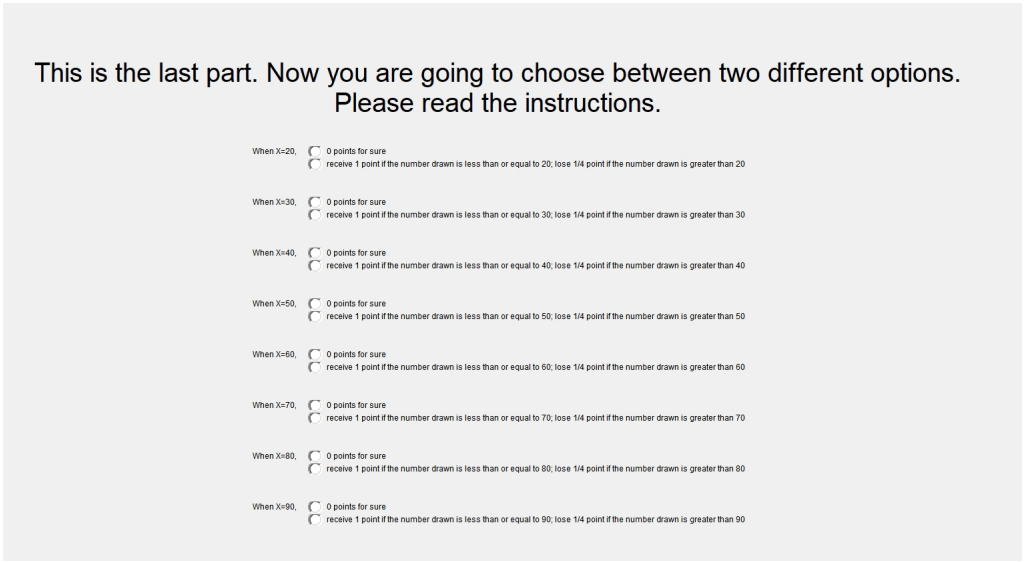
You are endowed with 5 points at the beginning of this part. In addition, the computer will randomly choose one of the choices you made in this part to determine your points.

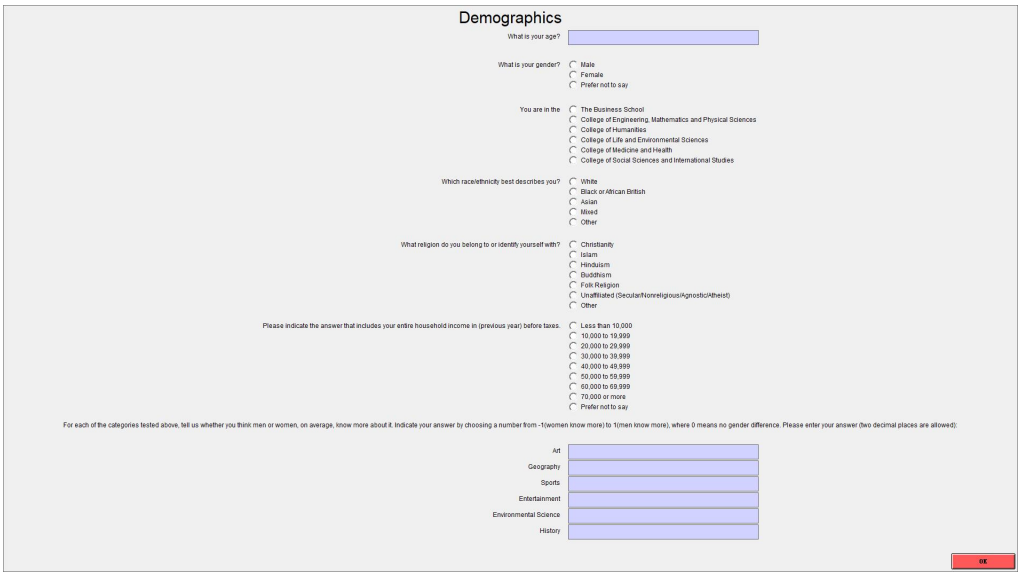
If you have any questions, please raise your hand. If you are ready, you may start now.

Questionnaire: We asked participants for demographic information and their gender stereotypes.

The gender stereotype question - For each of the categories tested above, tell us whether you think men or women, on average, know more about it. Indicate your answer by choosing a number from -1 (women know more) to 1(men know more), where 0 means no gender difference. Please enter your answer (two decimal places

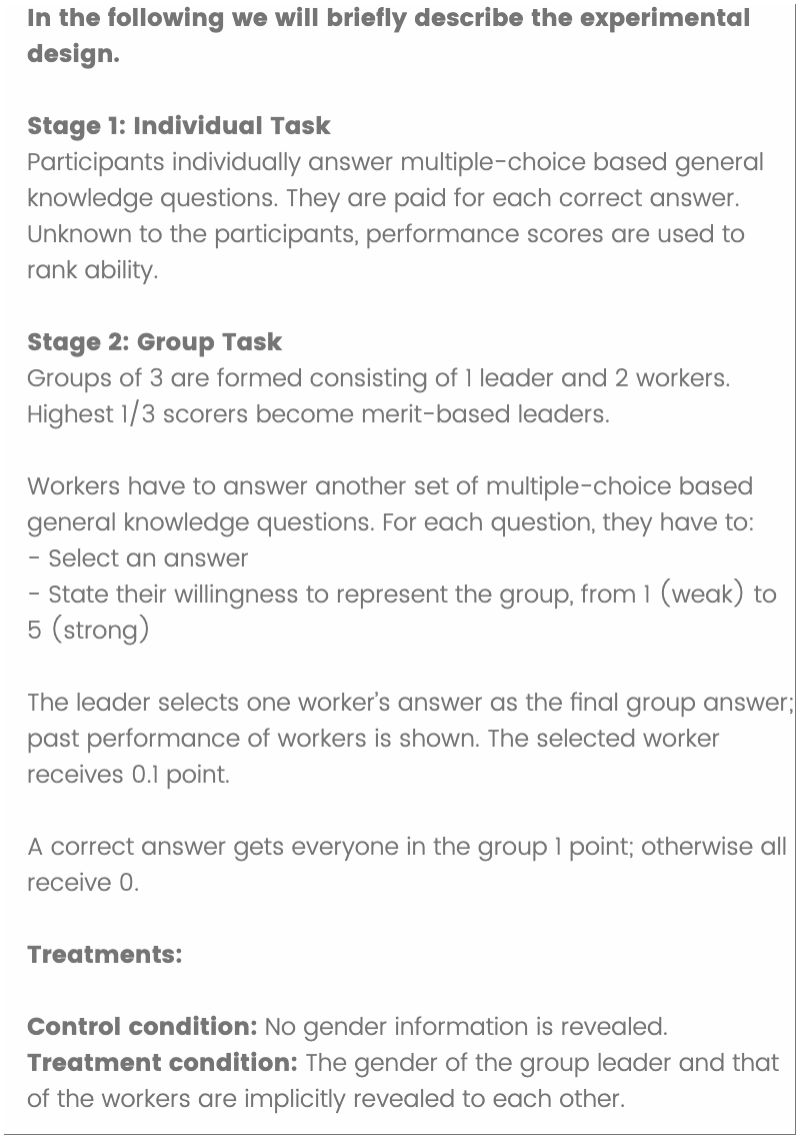
are allowed).


This is the last part. Now you are going to choose between two different options.
Please read the instructions.
When X=20,
0 points for sure
receive 1 point if the number drawn is less than or equal to 20; lose 1/4 point if the number drawn is greater than 20
When X=30,
0 points for sure
receive 1 point if the number drawn is less than or equal to 30; lose 1/4 point if the number drawn is greater than 30
When X=40,
0 points for sure
receive 1 point if the number drawn is less than or equal to 40; lose 1/4 point if the number drawn is greater than 40
When X=50,
0 points for sure
receive 1 point if the number drawn is less than or equal to 50; lose 1/4 point if the number drawn is greater than 50
When X=60,
0 points for sure
receive 1 point if the number drawn is less than or equal to 60; lose 1/4 point if the number drawn is greater than 60
When X=70,
0 points for sure
receive 1 point if the number drawn is less than or equal to 70; lose 1/4 point if the number drawn is greater than 70
When X=80,
0 points for sure
receive 1 point if the number drawn is less than or equal to 80; lose 1/4 point if the number drawn is greater than 80
When X=90,
0 points for sure
receive 1 point if the number drawn is less than or equal to 90; lose 1/4 point if the number drawn is greater than 90


Demographics
What is your age?
What is your gender?
Male
Female
Prefer not to say
You are in the
The Business School
College of Engineering, Mathematics and Physical Sciences
College of Humanities
College of Life and Environmental Sciences
College of Medicine and Health
College of Social Sciences and International Studies
Which race/ethnicity best describes you?
White
Black or African British
Asian
Mixed
Other
What religion do you belong to or identify yourself with?
Christianity
Islam
Hinduism
Buddhism
Folk Religion
Unaffiliated (Secular/Nonreligious/Agnostic/Atheist)
Other
Please indicate the answer that includes your entire household income in (previous year) before taxes.
Less than 10,000
10,000 to 19,999
20,000 to 29,999
30,000 to 39,999
40,000 to 49,999
50,000 to 59,999
60,000 to 69,999
70,000 or more
Prefer not to say
For each of the categories tested above, tell us whether you think men or women, on average, know more about it. Indicate your answer by choosing a number from -1(women know more) to 1(men know more), where 0 means no gender difference. Please enter your answer (two decimal places are allowed):
Art
Geography
Sports
Entertainment
Environmental Science
History
OK

6.3 Expert prediction task

The screenshots below show how we presented the experimental design to experts and asked for predictions in an experimental workshop in November 2025.



Your prediction:

What do you think is the percentage of correct answers (accuracy) of workers who state a willingness of 5 to represent their group under female leadership?

For each closest prediction, there is a £5 cash bonus.

0 10 20 30 40 50 60 70 80 90 100

Male worker's accuracy in %

Female worker's accuracy in %

What do you think is the percentage of correct answers (accuracy) of workers who state a willingness of 5 to represent their group under male leadership?

For each closest prediction, there is a £5 cash bonus.

0 10 20 30 40 50 60 70 80 90 100

Male worker's accuracy in %

Female worker's accuracy in %